Biomedical Text Mining for Disease Gene Discovery

Sarah ElShal[1,2]✉, Jesse Davis[3], Yves Moreau[1,2]

[1]Dept. of Electrical Engineering (ESAT-SCD) - K.U.Leuven, Leuven, Belgium
[2]IBBT-KULeuven Future Health Department, Leuven, Belgium
[3]Dept. of Computer Science- K.U.Leuven, Leuven, Belgium

Motivation and Objectives

Because of the amount of electronic literature now available, it is challenging for biologists to search biomedical corpuses for any kind of desired information beyond simple text retrieval. Several tools have been developed to make text mining easier for them. Some of these tools focus on extracting biomedical terms; such as protein names and biological processes, given any input text. The tools COREMINE Medical (http://www.coremine.com, last accessed on 25 September 2012) and GoPubMed: http://www.gopubmed.com, last accessed on 25 September 2012) are just two examples. Other tools apply rule-based strategies to relate biomedical concepts to each other. E.g., BITOLA (http://ibmi.mf.uni-lj.si/bitola/ last accessed on 25t September 2012) (Hristovski et al., 2005).

We have been developing a methodology and tool to discover genes implicated in any given disease or disorder. In fact, our tool takes from the user any free text query as an input and attempts to identify those genes most strongly linked to the query. As an output, the tool returns an ordered list of the best genes matching the query. The core work of our tool is based on text mining. Basically, each gene is linked to a profile that contains the biological terms that are most significant for it. Similarly, we link the input query to a corresponding keyword profile. The genes appearing at the top of the output list are the ones whose profiles are highly similar to that of the input query.

Methods

The text mining strategies we use in our work are applied to the biomedical abstracts published in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/, last accessed on 25 September 2012). We divide our work into two phases: a background phase, and a live phase. In the background phase, we collect all the abstracts annotated to every gene described in Entrez Gene (http://www.ncbi.nlm.nih.gov/gene/, last accessed on 25 September 2012). For this, we use GeneRIF http://www.ncbi.nlm.nih.gov/gene/about-generif last accessed on 25 September 2012), which provides functional annotation between genes and PubMed references. Afterwards, we index all the referenced abstracts via MetaMap (http://metamap.nlm.nih.gov/, last accessed on 25 September 2012) (Aronson, 2001), which maps the given biological text to the Unified Medical Language System (UMLS) Metathesaurus (http://www.nlm.nih.gov/research/umls/, last accessed on 25 September 2012) (Bodenreider, 2004). Thus for each gene, we could maintain a list of UMLS biomedical terms that functionally-describe it. We call this list a gene keyword profile. Then for each gene, we build another weighted profile in the form of a vector of Term Frequency-Inverse Document Frequency (TF-IDF). For a given gene, each entry in the vector measures how relevant a specific UMLS term is to the gene. We refer to the whole set of gene vectors as the "reference matrix". An example of this reference matrix is shown below in Table 1.

In the live phase of our work, we take a free text query as an input from the user (e.g., sleep disorders). Then, we use the E-utilities from PubMed to retrieve the corresponding abstracts that are relevant to the user query. And as we did with the genes in the background phase, we generate a keyword profile for the query and consequently a corresponding TF-IDF vector. Finally we match this query vector against all the gene vector entries in the reference matrix. Each match corresponds to a score that is calculated via a dot-product. The higher the matching score of a given gene vector entry, the more probably the gene relates to the user query. We also take into account the frequency of citation of a given gene. So genes appearing early in the ordered output list, do not only share the most similar profiles with the user query, but they are also cited by the highest number of references. Besides, we consider the fraction of common references between the query and the candidate genes as an additional scoring factor. As the number of com-

Table1: An example of the Reference Matrix. The heading row corresponds to a set of different UMLS terms (DNA-binding, cancers, tumours, …, diabetes, and peptides). The heading column corresponds to two gene examples (Breast Cancer Type 1 (BRCA1), and Insulin (INS)). The numbers in each row (vector) correspond to the TF-IDF values of each UMLS term given the heading gene. For example, we observe that for BRCA1, the terms "Cancers" and "Tumours" have high TF-IDF values. This is related to the fact that they have high frequency of occurrence in the abstract texts annotated to BRCA1. Besides, we also observe that "DNA-Binding", "Diabetes", and "Peptides" have low TF-IDF values since they are not that frequent in the annotated text.

	DNA-Binding	Cancers	Tumors	…	Diabetes	Peptides
BRCA1	0.0	10.3	9.8	…	2.3	0.0
INS	0.0	3.7	0.0	…	10.5	9.3

mon references increases, the matching score of the candidate gene also increases.

Results and Discussion

The evaluation of the results is still ongoing. To validate the quality of our results, we use as a benchmark the phenotype-gene annotation provided by the Human-Phenotype-Ontology (HPO http://human-phenotype-ontology.org/, last accessed on 25 September 2012) (Robinson and Mundlos, 2010). For every phenotype in this annotation, a set of linked genes are recorded. The links are provided based on the information about the phenotypes of a given syndrome, and the genes known to cause that syndrome. Hence in our tool, we use each phenotype in the annotation file as a separate free text input. Then for each gene output list, we measure the percentage of recall against the HPO annotation.

To assess the power of our tool, we use some general biomedical search systems as a baseline (e.g., Gene Ontology http://www.geneontology.org/, last accessed on 25 September 2012). We are expecting our tool to perform better. That is because such general systems rely on clear evidence to associate a gene product with a given query (e.g., inference from experiments or by curators), while our tool digs deeper in all the published literature as discussed in the Methods section.

References

1. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp: 17.
2. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 32: D267. doi:10.1093/nar/gkh061
3. Hristovski D, Peterlin B, et al. (2005) Using literature-based discovery to identify disease candidate genes. Int J Med Inform 74: 289.
4. Robinson PN, Mundlos S (2010) The Human Phenotype Ontology. Clin Genet 77: 525